Edge AI Over Wireless:

Present & Future

Mehdi Bennis

Associate Professor

Head of ICON, IEEE Fellow

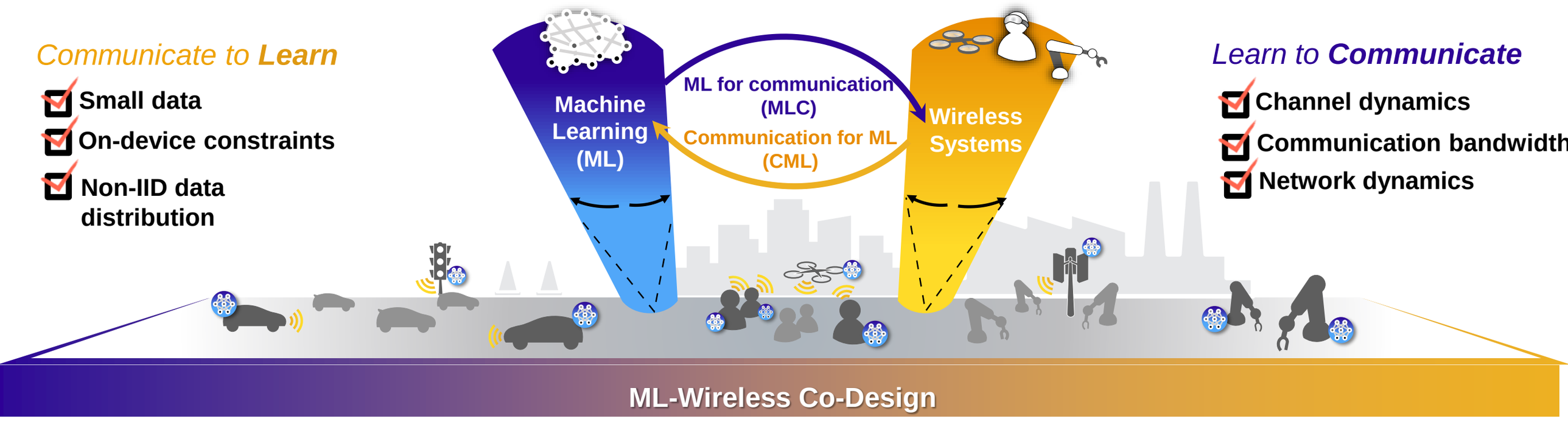
University of Oulu

OPEN DAY CTM 2021

Proliferation of intelligent devices & mission-critical applications at the network edge cannot be operated with centralized and best-effort ML

*Communicate to **Learn***

- ☑ Small data
- ☑ On-device constraints
- ☑ Non-IID data distribution

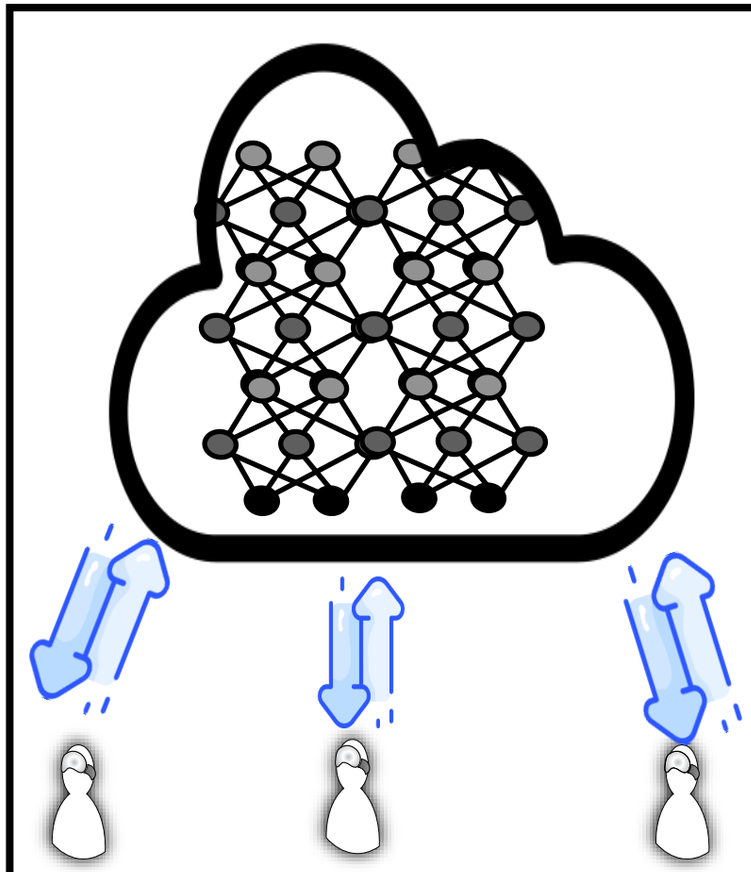


*Learn to **Communicate***

- ☑ Channel dynamics
- ☑ Communication bandwidth
- ☑ Network dynamics

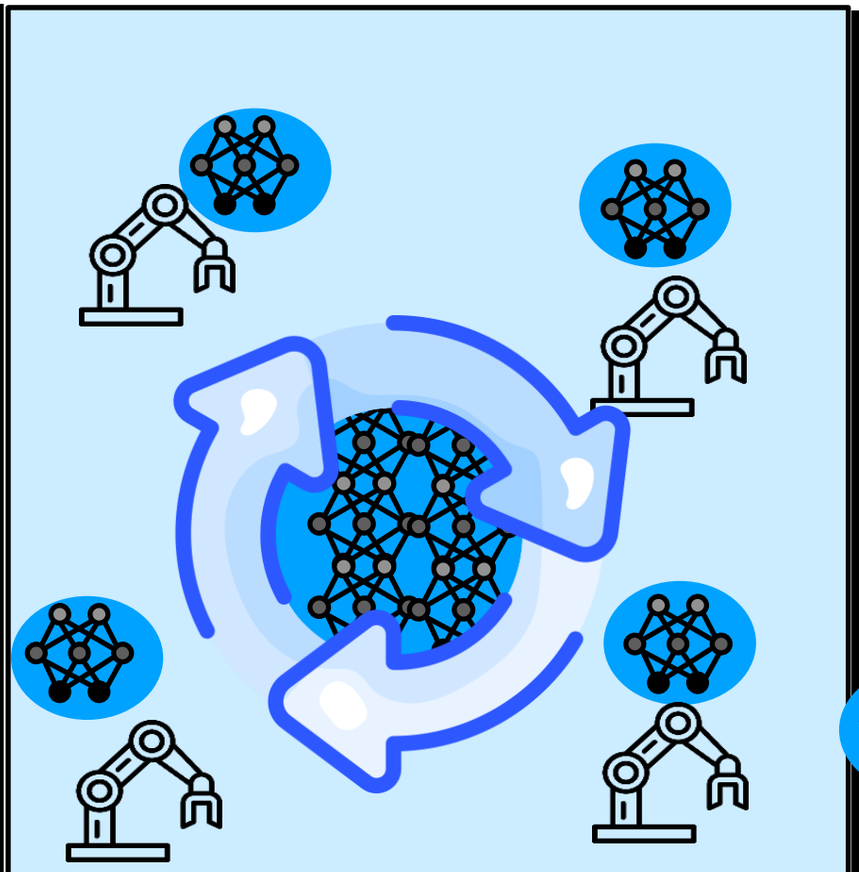
Communication-efficient, low-latency, reliable and scalable (i) training; (ii) inference; (iii) control

Centralized to Federated and Swarm/Distributed ML



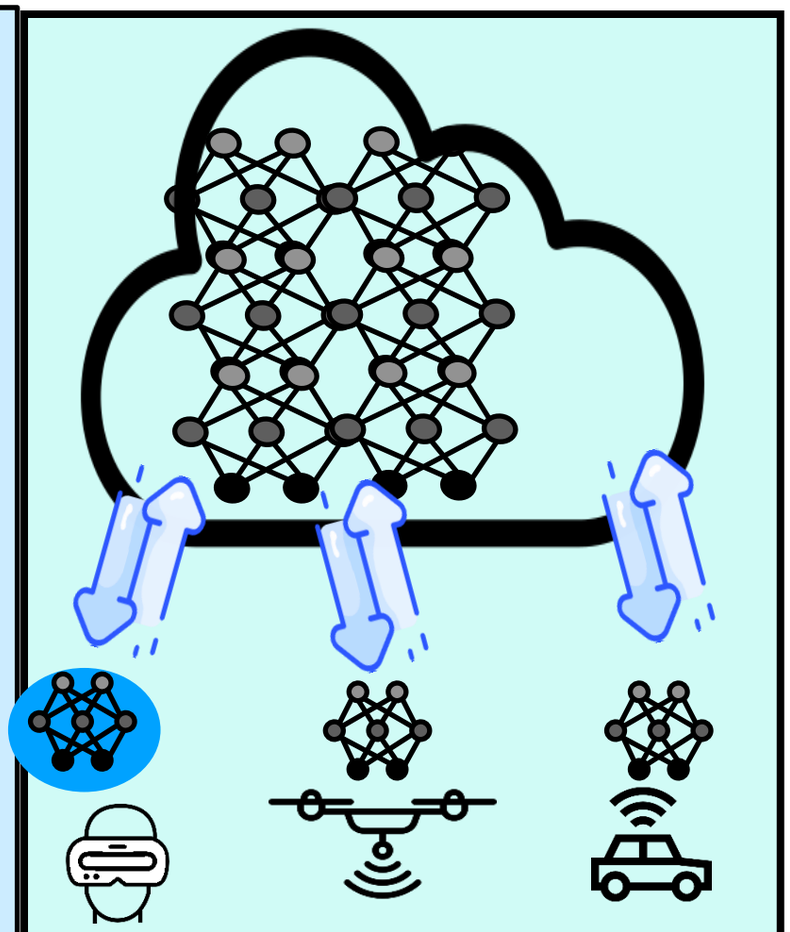
- No Privacy ✗
- High latency ✗
- Bandwidth inefficient ✗

Classical AI



- Collective intelligence / Fully autonomous ✓
- Scalable ✓
- Privacy-preserving ✓

Swarm AI



- Bandwidth efficient ✓
- Use cloud but smartly ✓
- Privacy-preserving ✓

Federated AI

NOKIA HUAWEI SAMSUNG intel PHILIPS

Big Picture

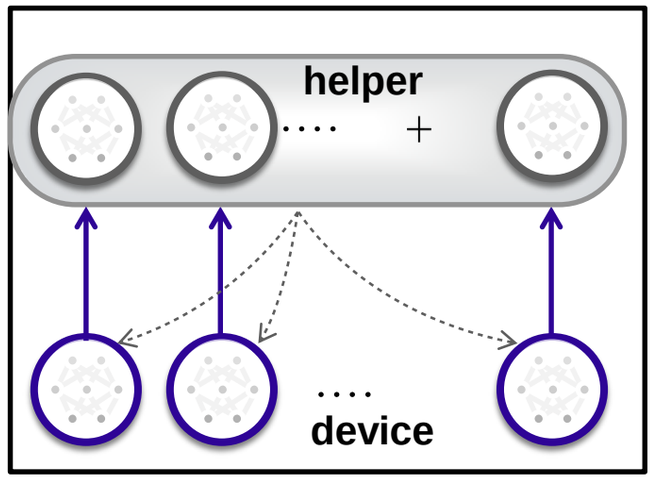
Sol.

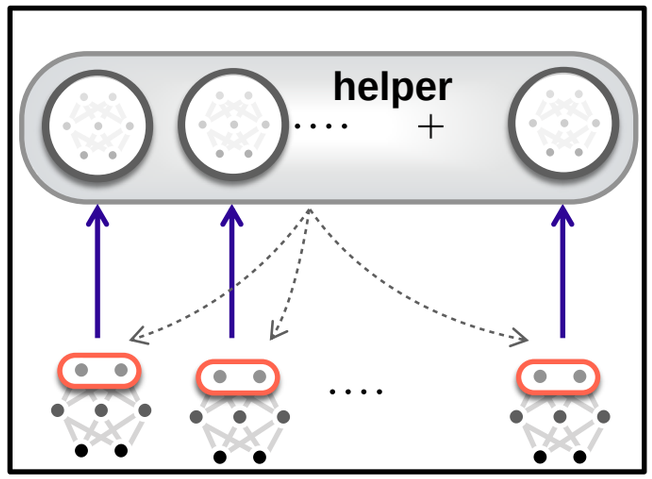
- **Federated Learning**
- **Group ADMM**
- Blockchained FL

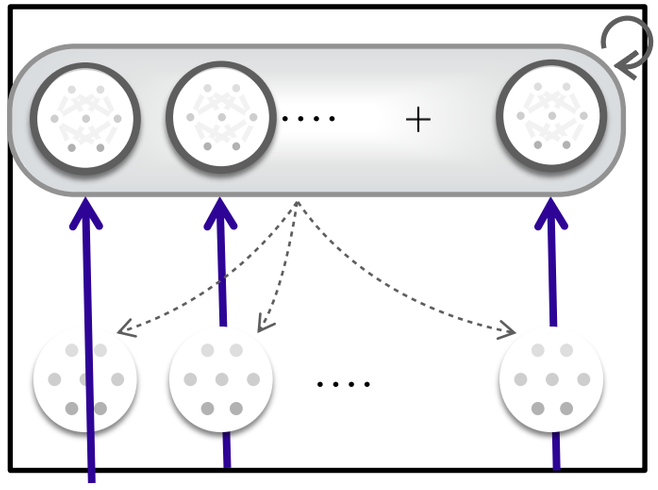
- **Federated Distillation**
- FL after Distillation

- Data Summarization
- Federated Augmentation

Type



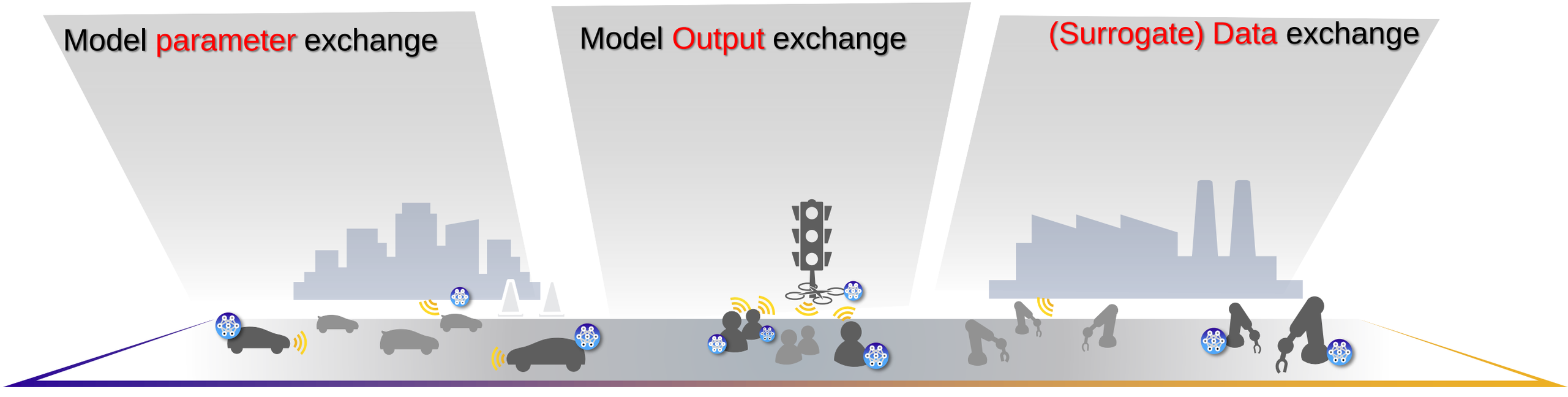




Model parameter exchange

Model Output exchange

(Surrogate) Data exchange



ExtFL = FL + **Extreme Value Theory**

Extreme Queue Length FL for Vehicular URLLC Power Control

Problem. Minimize vehicular user equipment (VUE)'s avg. uplink power, subject to each VUE's **queue length reliability** $\Pr(Q > q_{\text{th}}) \leq \epsilon$

- Following **extreme value theory (EVT)**, an extremely large queue length is characterized by the shape and scale parameters of the generalized Pareto distribution (GPD)
- Utilizing **FL with EVT (ExtFL)**, vehicular user equipments collectively predict the GPD parameters
- ExtFL **reduces communication overhead** while achieving the same queue length reliability, compared to a centralized direct queue length distribution exchanging baseline (CEN)

Sample

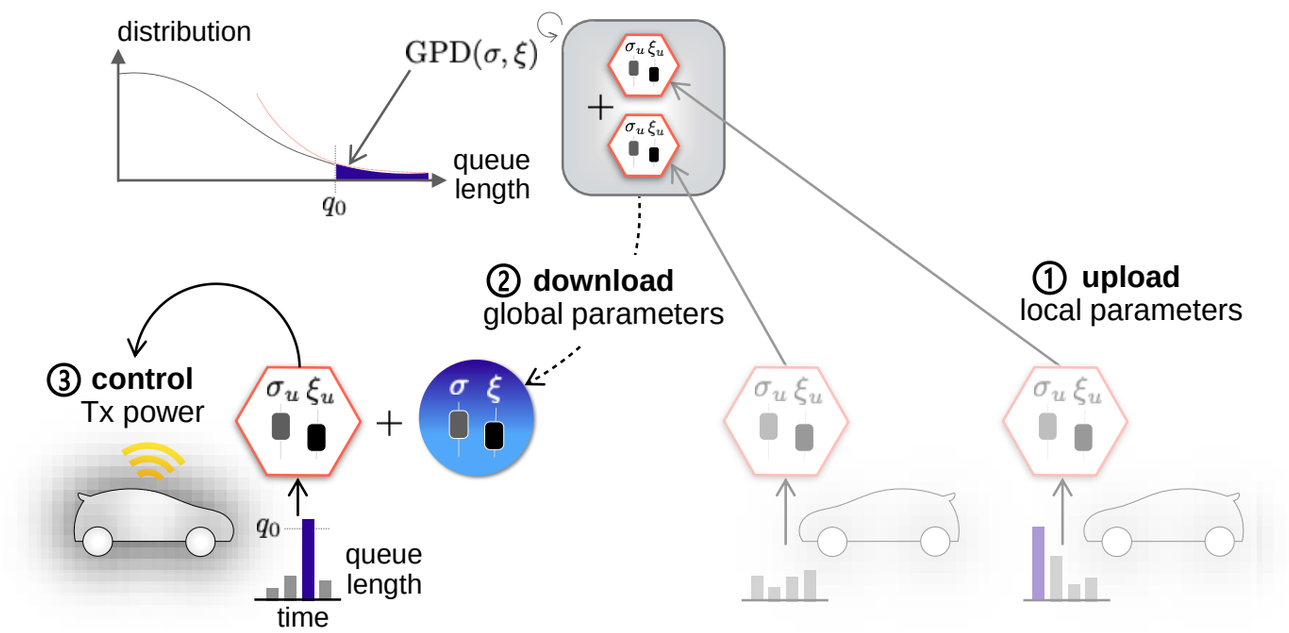
Empirical distribution

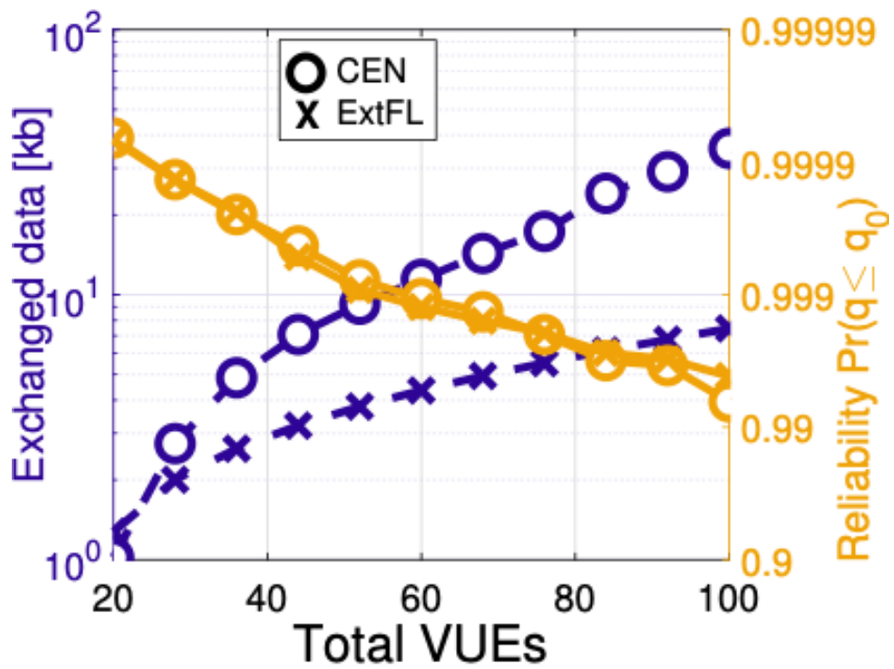
$$\min_{\sigma,\xi} -\frac{1}{N}\sum_{n}\log\left(G(x_n,\sigma,\xi)\right)$$

Scale parameter

Shape parameter

Extreme Value Theoretic FL (ExtFL)





S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "**Distributed Federated Learning for Ultra-Reliable Low-Latency Vehicular Communications**," *TCOM'20*

Beyond Federated (server-based) Learning

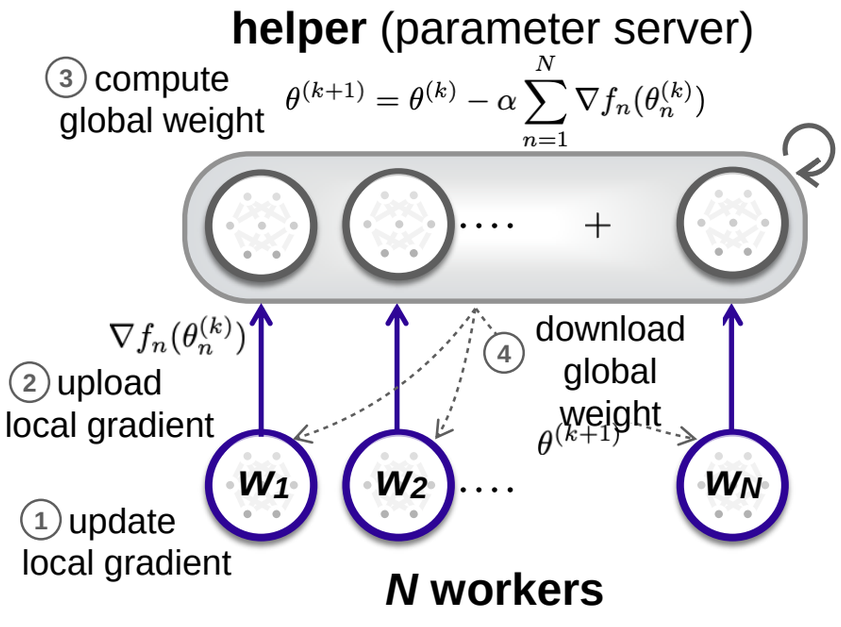
Group ADMM (without any central entity)

- Idea. Exploiting ADMM for faster training convergence without any central entity

1) **Head** devices update **primal** variables (weights) in parallel
2) Each **head** device transmits the weights to its (two) **neighboring tail** devices
3) **Tail** devices update **primal** variables in parallel
4) Each **tail** device transmits the weights to its **neighboring head** devices
5) Each device updates its dual variable

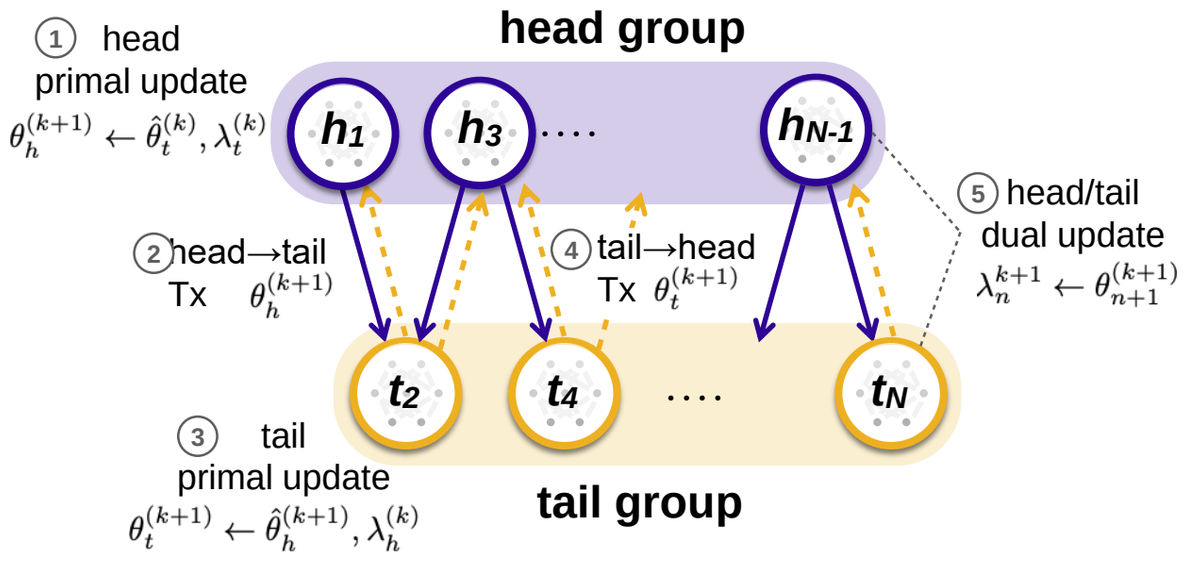
FL (GD based)

$$\underset{\{\theta_n\}}{\text{Minimize}} \quad \sum_{n=1}^{N} f_n(\theta_n)$$
$$\text{s.t.} \quad \theta_n = \theta \quad \forall n$$



GADMM

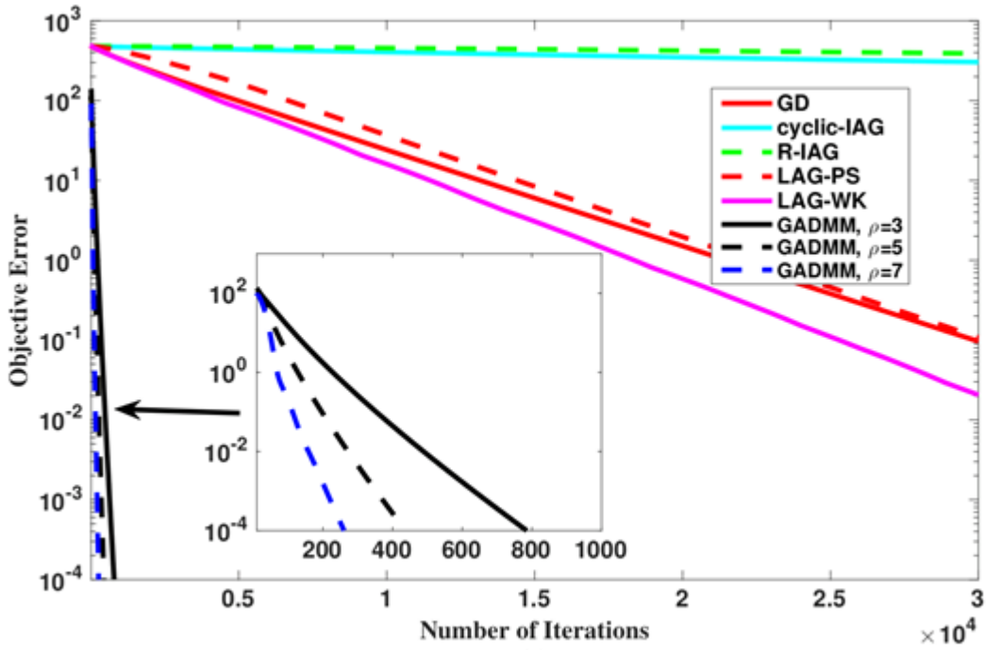
$$\underset{\{\theta_n\}}{\text{Minimize}} \quad \sum_{n=1}^{N} f_n(\theta_n)$$
$$\text{s.t.} \quad \theta_n = \theta_{n+1} \quad \forall n$$

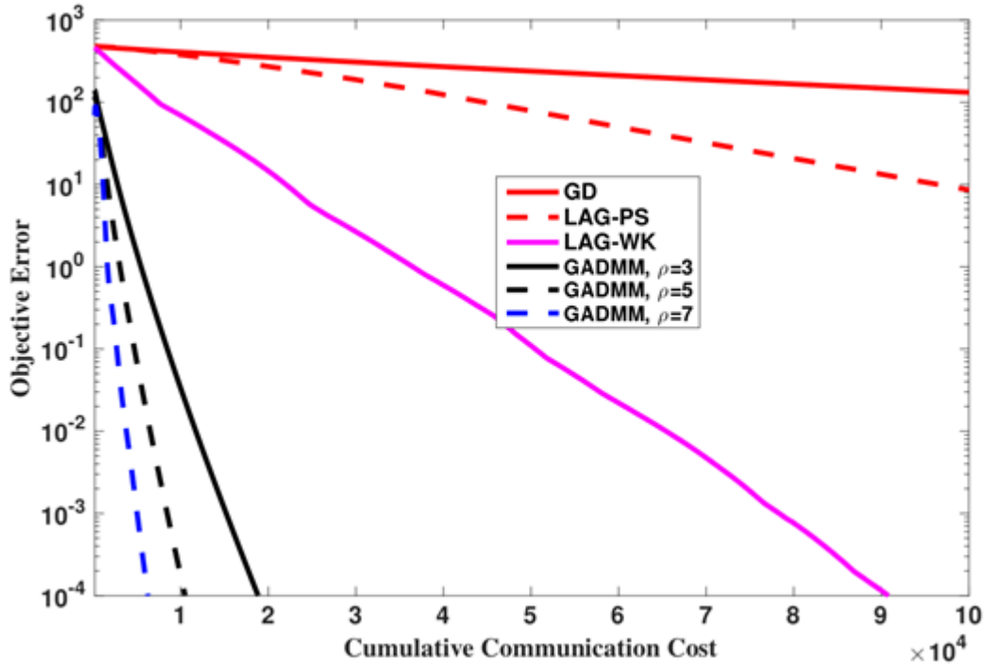


A. Elgabli, J. Park, A. S. Bedi, V. Aggarwal, and M. Bennis, "**GADMM: Fast and Communication Efficient Distributed Machine Learning Framework**," *JMLR20*

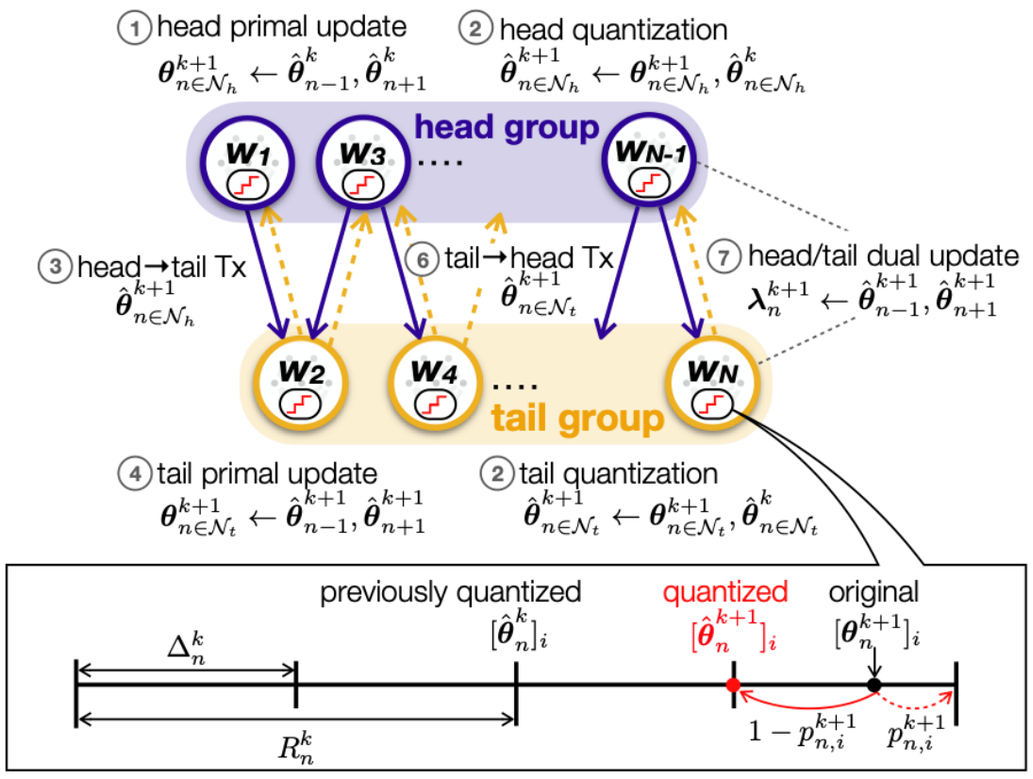
GADMM

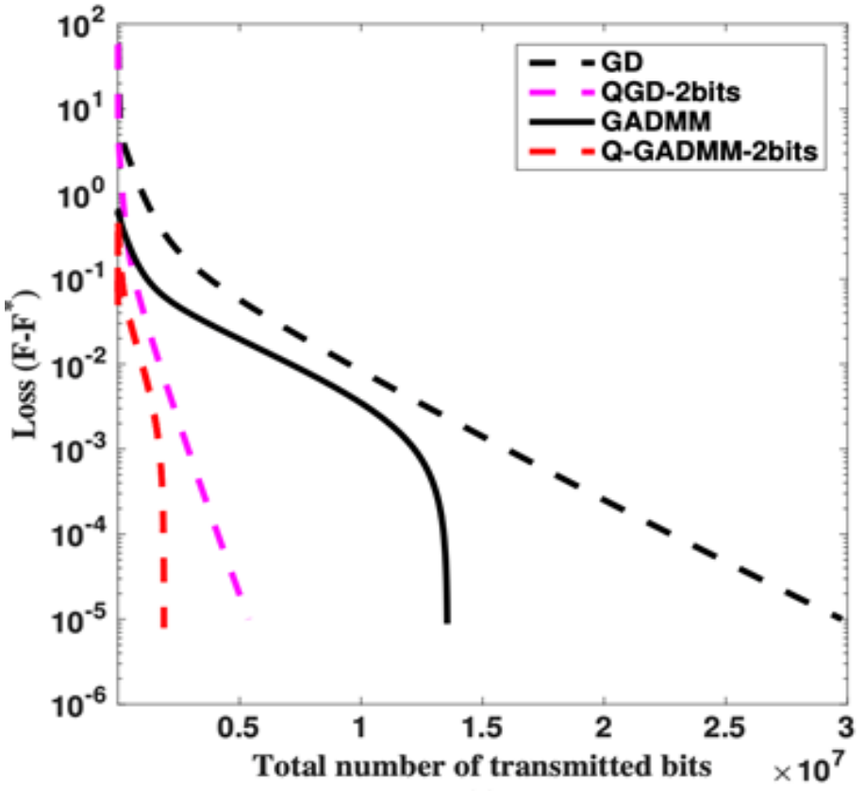
GADMM, Linear Regression





Quantization

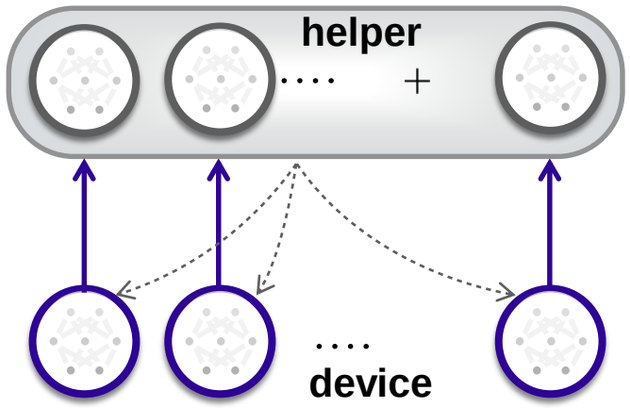




Wireless *Analog* FL

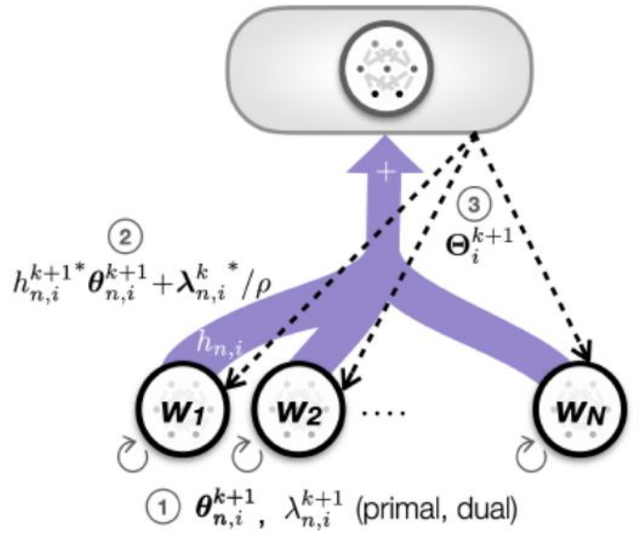
Digital FL

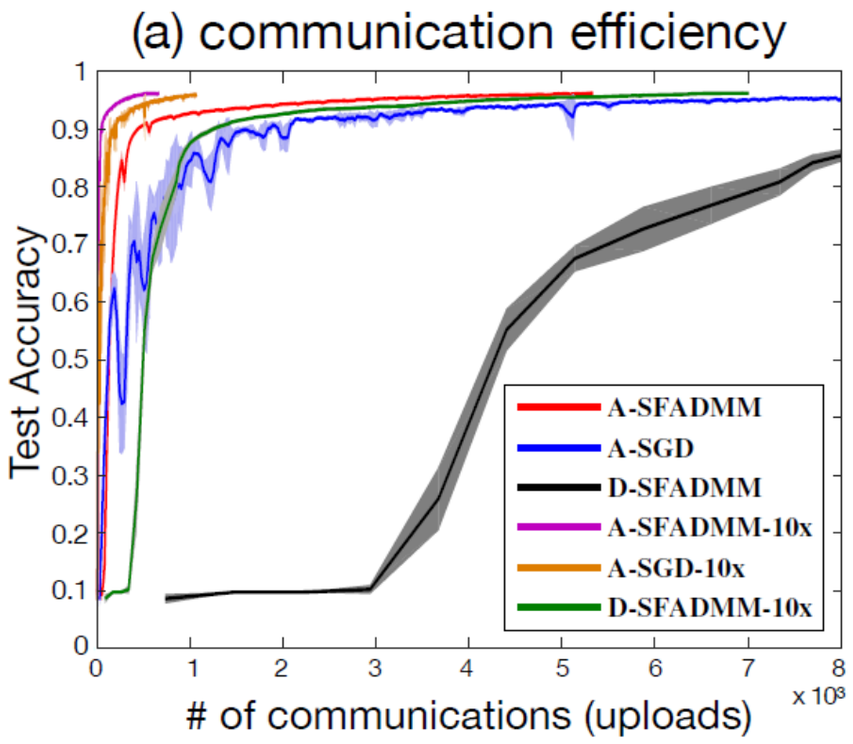
- **Orthogonal** transmission
- Does not scale across no. of workers and model size ✗

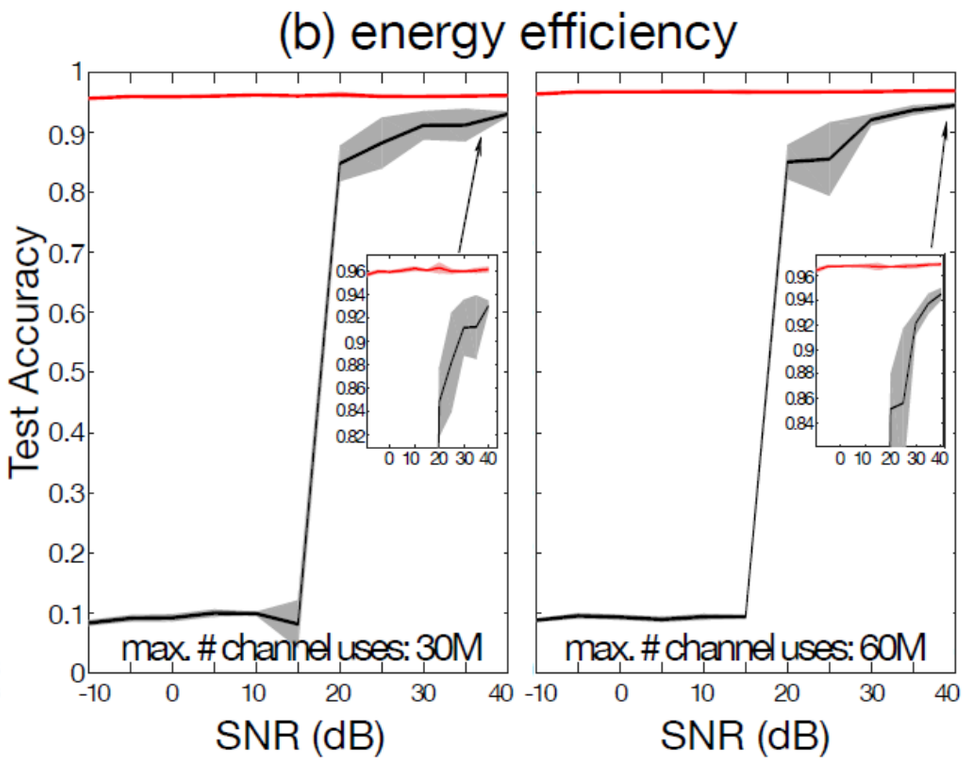


Analog FL

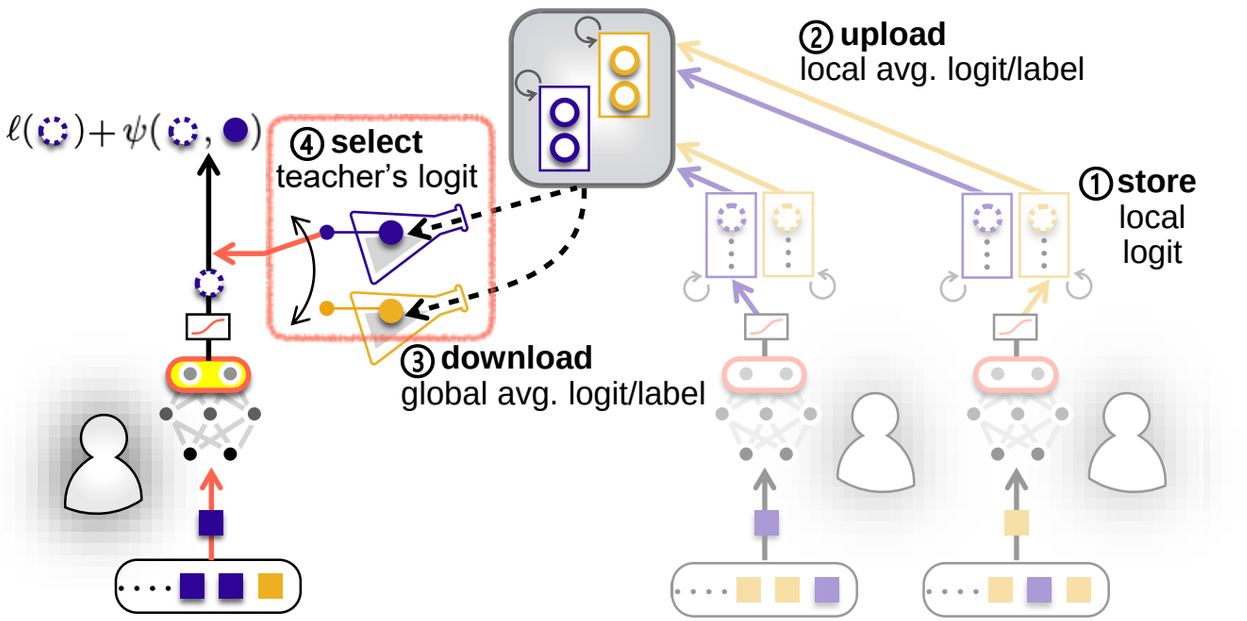
- Non-orthogonal transmission and channel **superposition**
- Scaling across no. of workers and model size ✓
- Fast & communication-efficient

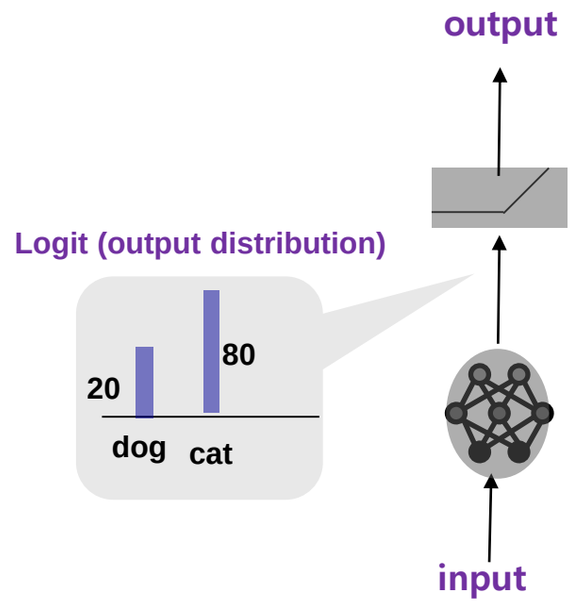






Federated Distillation (FD)

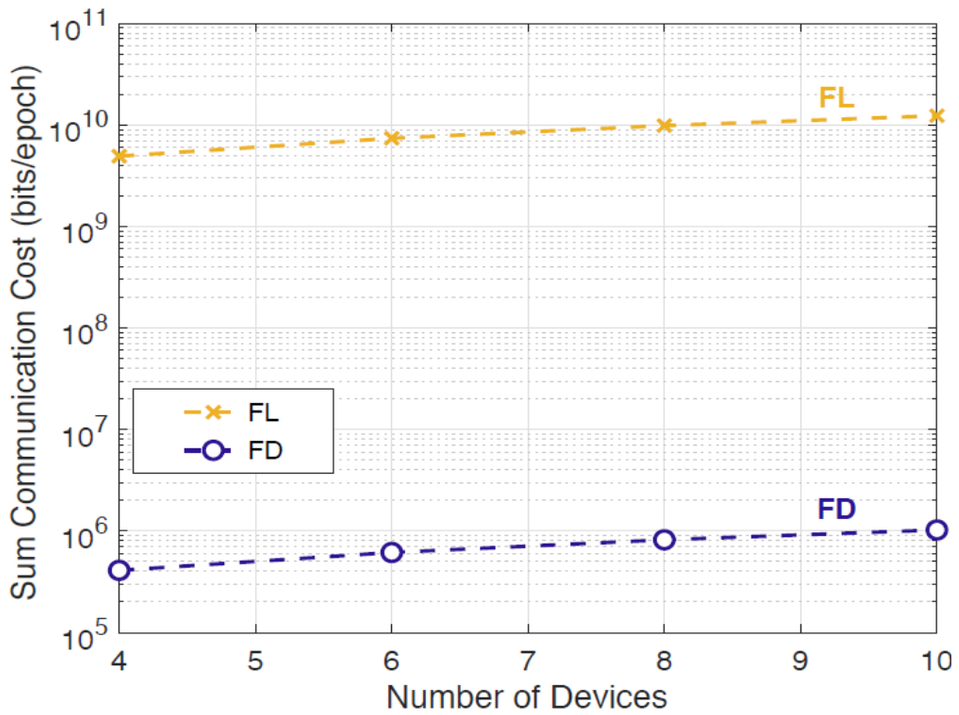


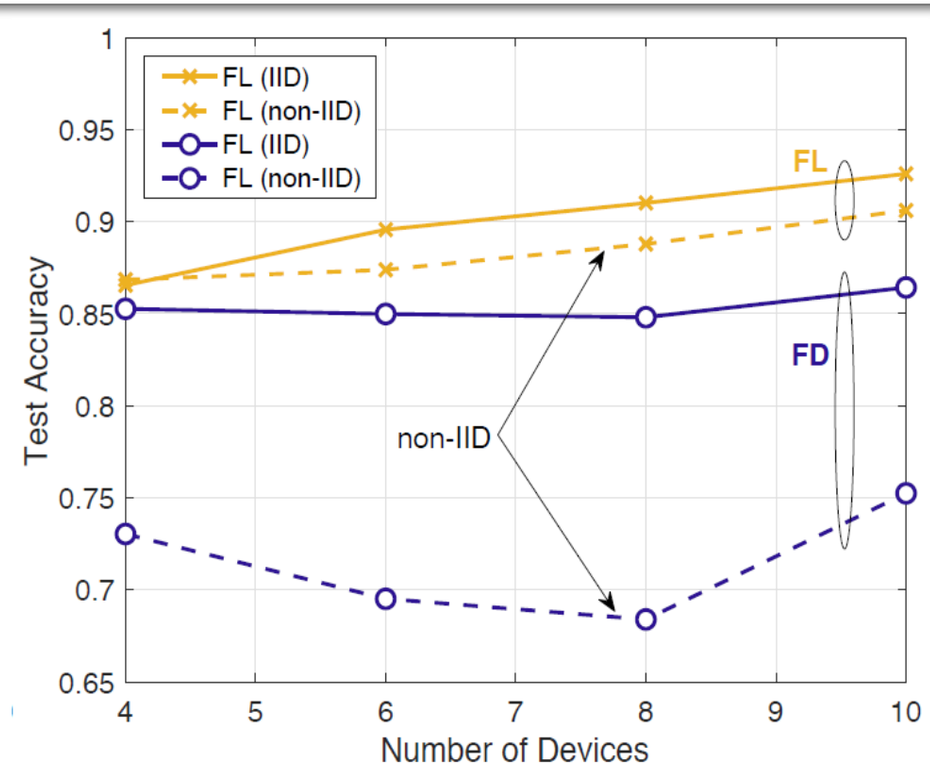


$$w_{k+1}^{(i)} = \begin{cases} w_k^{(i)} - \eta \left(\nabla \ell(w_k^{(i)}) + \psi(F_{k,\ell}^{(i)}, \breve{F}_{k,\ell}^{(i)}) \right) & \text{if } k \bmod \tau = 0, \\ w_k^{(i)} - \eta \nabla \ell(w_k^{(i)}) & \text{otherwise} \end{cases}$$

teacher's logit = global avg. logit/label $\breve{F}_{k,\ell}^{(i)} = \sum_{j \neq i} \hat{F}_{k,\ell}^{(j)} / M$

local avg. logit/label $\hat{F}_{k,\ell}^{(i)} = \sum_{k} F_{k,\ell}^{(j)} / \tau$

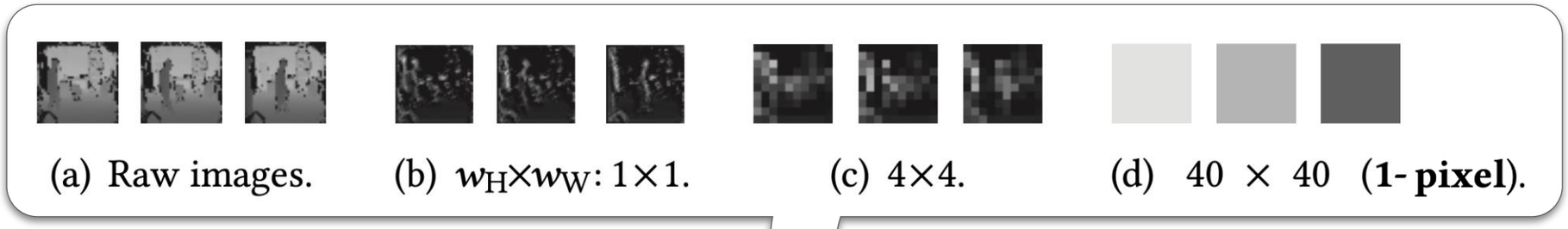


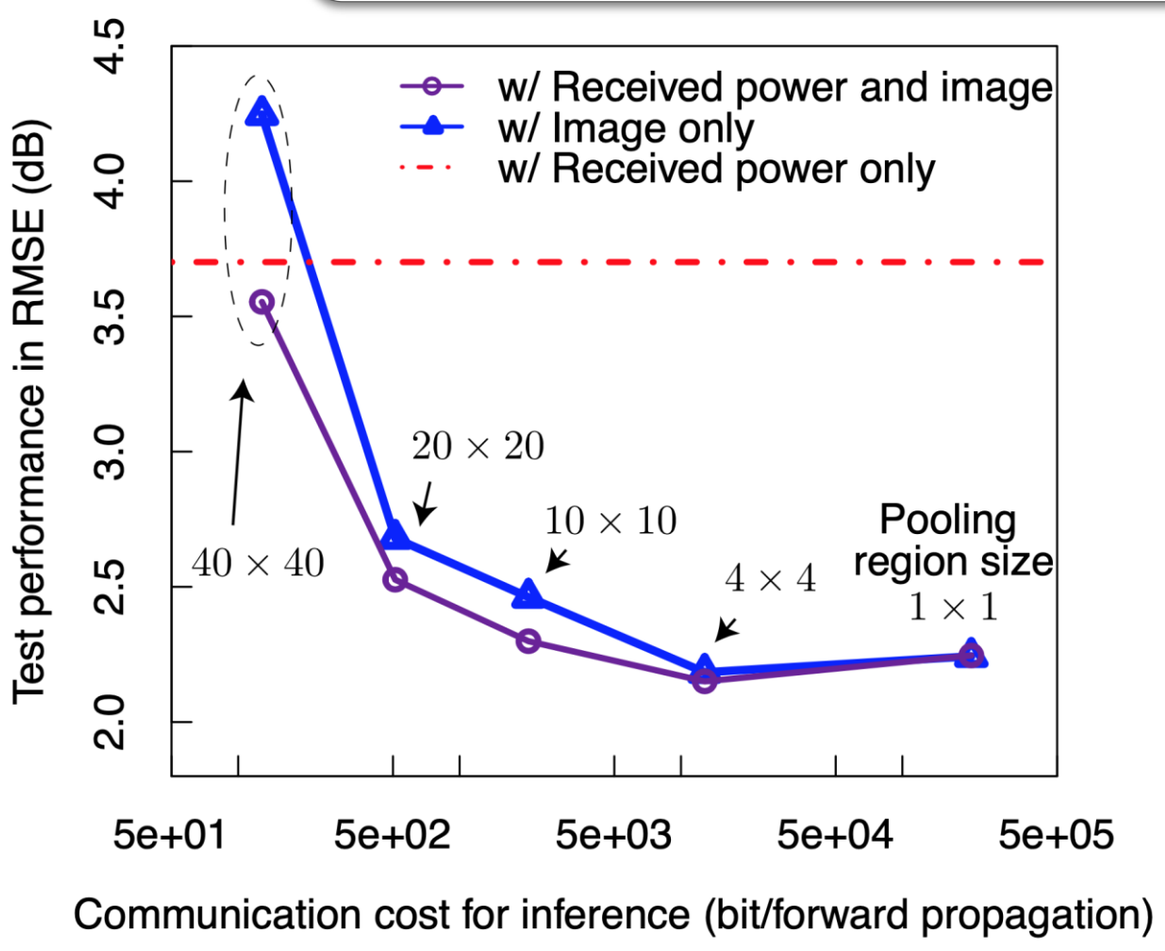


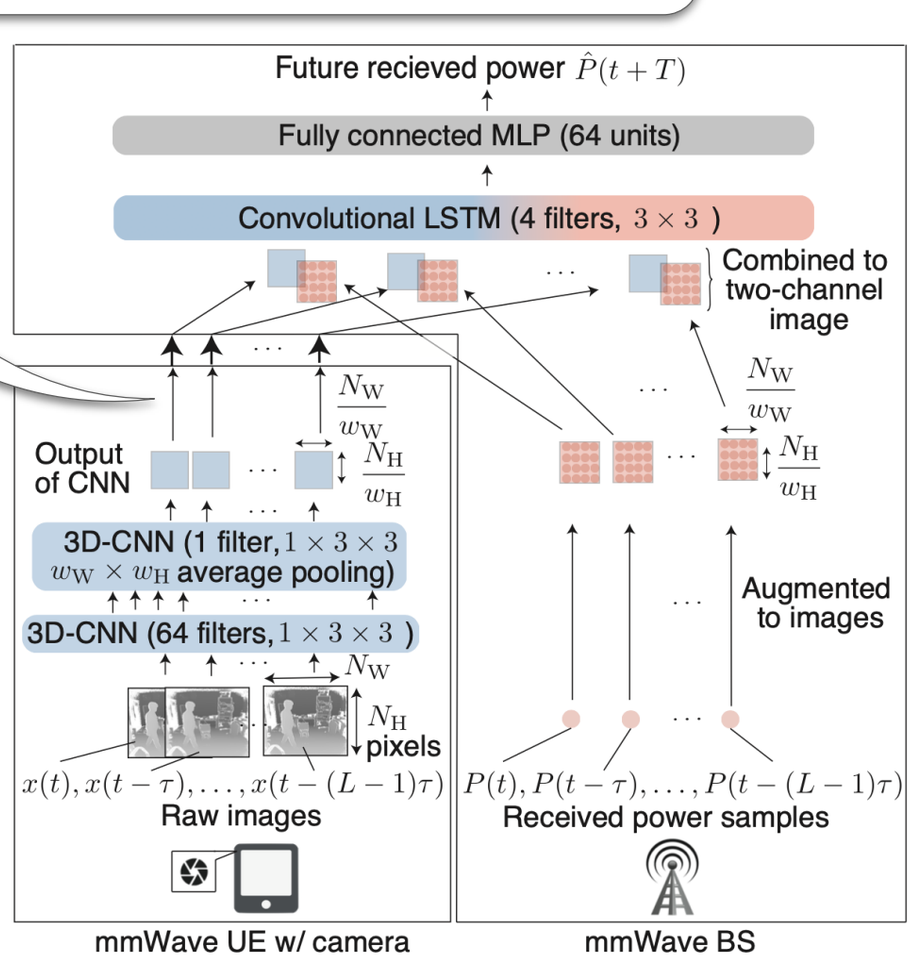
E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "**Federated Distillation and Augmentation under Non-IID Private Data**," *NeurIPS 2018 MLPCD*

Split Learning for mmWave Channel Blockage Prediction

- Images are processed by a CNN whose output is **transmitted** to the BS
- BS's LSTM layers accept the concatenated CNN output and RF signals
- Idea: The higher **CNN pooling dimension** w_W x w_H, the lower communication payload sizes
- **Accuracy vs. communication**: Accuracy is optimized at 4x4, while the payload is minimized at 40x40
- Even **one-pixel image** (40x40 pooling dimension) improves the performance







What’s Next?

Limitations

- Obsession with accuracy
  - Energy Bill? Sustainability?
- Brittle, lacks robustness; Poor Generalization
- FL is the **first-step** towards **truly intelligent systems (6G)**

➨**Function approximators (curve fitting + learning *CORRELATIONS*).**

➨ **Lack extrapolation..**

Desiderata

1. **Function** of data
2. **Minimal** without compromising the **sufficient** effectiveness in the task
3. **Invariant**
4. **Disentangled**
5. **Causal** for extrapolating OOD data

Objective

Learning **Semantic** representations satisfying **D1-D5** for **X**

☑ Tx less data ☑ More reliable ☑ More energy-efficient ☑ Sample efficient ☑ Intelligent

Thank you!

https://sites.google.com/view/dr-mehdi-bennis/research/edge-ai?authuser=0